All You Ever Need to Know About Counterfactual Explanations

(and were afraid to ask...)

Fundamentals, Methods, & User Studies for XAI

André Artelt ~ Ulrike Kuhl ~ Mark T. Keane

André Artelt

Bielefeld University, Germany
University of Cyprus, Cyprus

Ulrike Kuhl

Faculty of Technology
Bielefeld University
Bielefeld, Germany

Mark T. Keane

School of Computer Science
University College Dublin
Dublin, Ireland

UNIVERSITÄT BIELEFELD
Faculty of Technology

University of Cyprus

UNIVERSITÄT BIELEFELD
Faculty of Technology

dataninja.nrw

Funded by the
Ministry of Culture and Science of the State of North Rhine-Westphalia

UCD DUBLIN
University College Dublin
Ireland's Global University

Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS

The Problem

ChatGPT OpenAI

https://youtu.be/eMx-2s7mZ24

AI is out there...being deployed...busily making mistakes !

Spot the Ball !

Spot the Head !

One Solution: Explain It !

- *Ante-Hoc*: Somehow show the workings of the machine (*simulatability*)
- *Post-Hoc:* Justification after-the-fact to *explain* workings (*evidence of what happened*)

Post Hoc: It's Wrong Because !

Test Image

Label: **Rifle**

Classification: **Shovel**

Explanation-by-Example

Label: **Shovel**

Classification: **Shovel**

Kenny, Delaney & Keane (2023) IJCAI-23

Post Hoc: To Audit a CNN !

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA-ENABLED INTELLIGENCE FOR DIGITAL HEALTH

Received September 14, 2019, accepted October 7, 2019, date of publication October 21, 2019, date of current version October 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2948430

Explainable Prediction of Chronic Renal Disease in the Colombian Population Using Neural Networks and Case-Based Reasoning

GABRIEL R. VÁSQUEZ-MORALES[1], SERGIO M. MARTÍNEZ-MONTERRUBIO[2], PABLO MORENO-GER[3], AND JUAN A. RECIO-GARCÍA[1]

[1]Office of Information and Communications Technology, Ministry of Health and Social Protection, Bogotá 110311, Colombia
[2]Group of Artificial Intelligence Applications, Department of Software Engineering and Artificial Intelligence, Faculty of Computer Science, Universidad Complutense de Madrid, Ciudad Universitaria, 28040 Madrid, Spain
[3]School of Engineering, Universidad Internacional de La Rioja (UNIR), 26006 Logroño, Spain

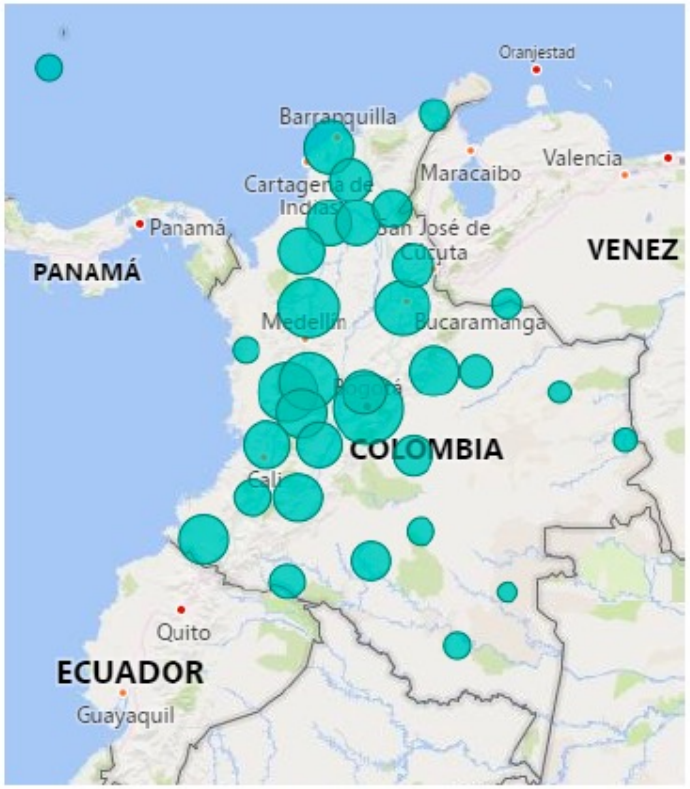
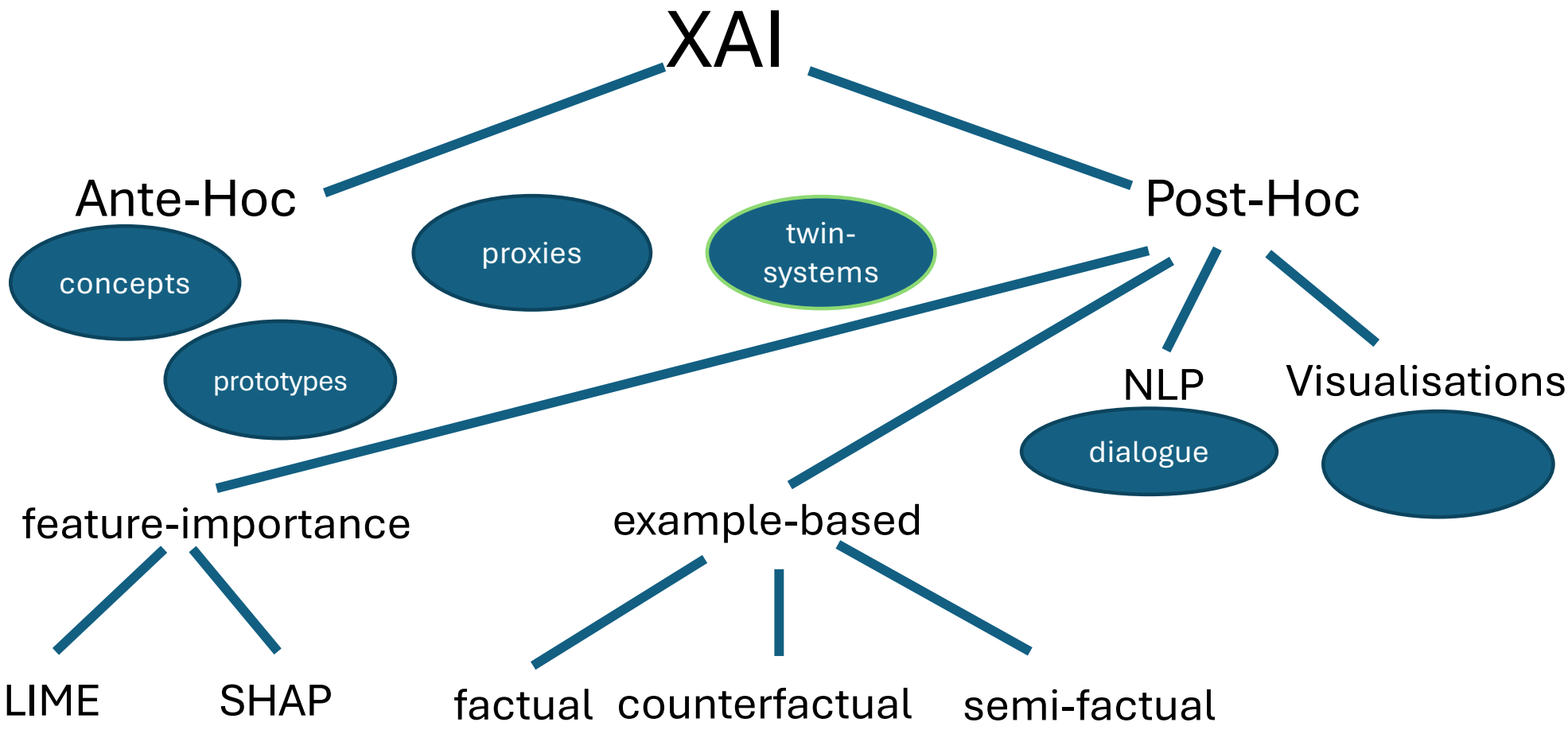
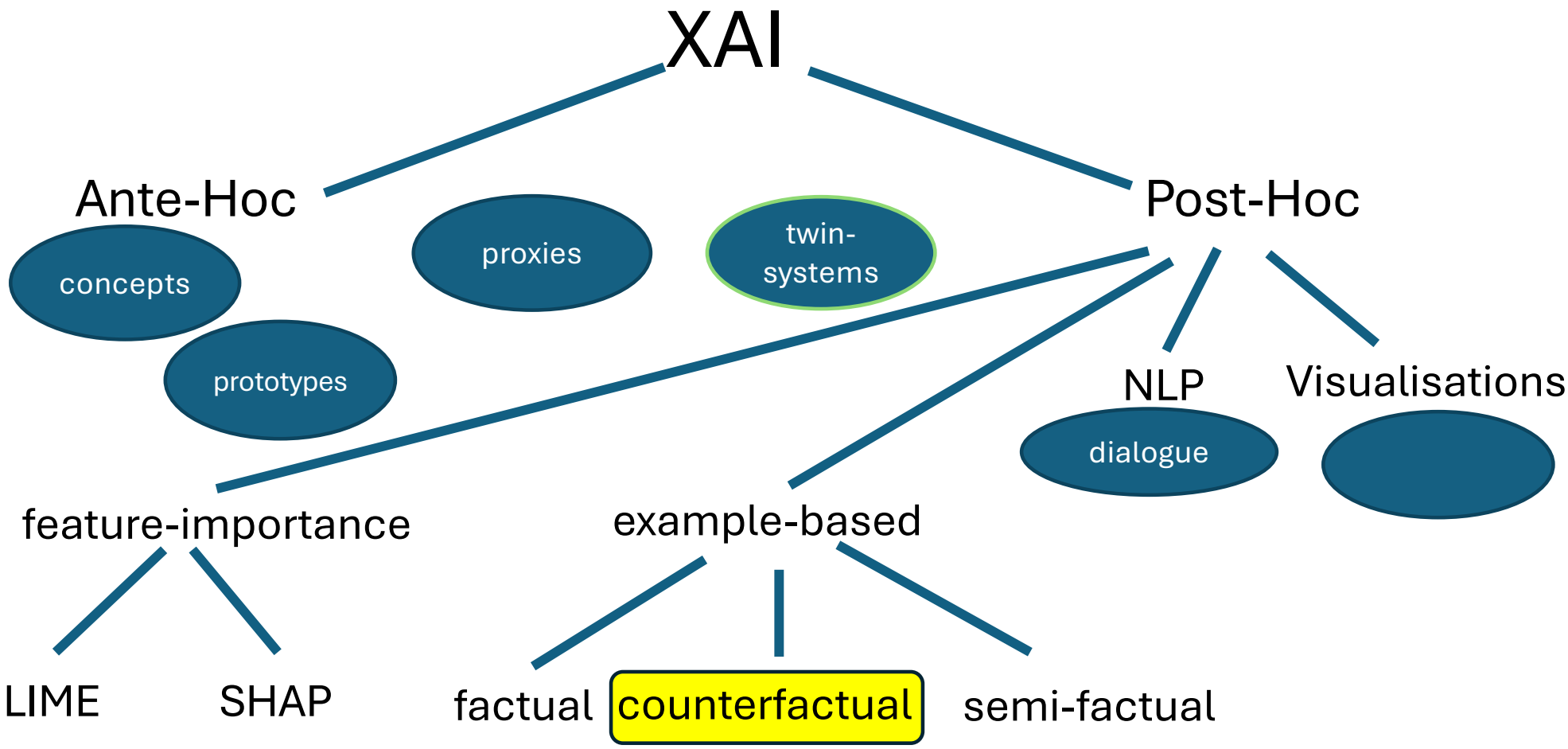


FIGURE 17. Map of coverage of the population at risk of developing CKD.

Explainable AI: Taxonomy



Explainable AI: Taxonomy



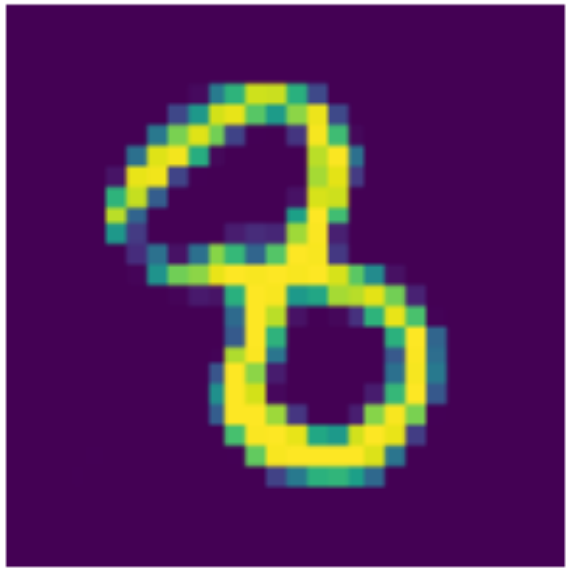
Counterfactual Explanations

Counterfactual Explanations

If the test image looked like this, I would have thought it was an "8".

Test Image

Label: 8
Prediction: 3



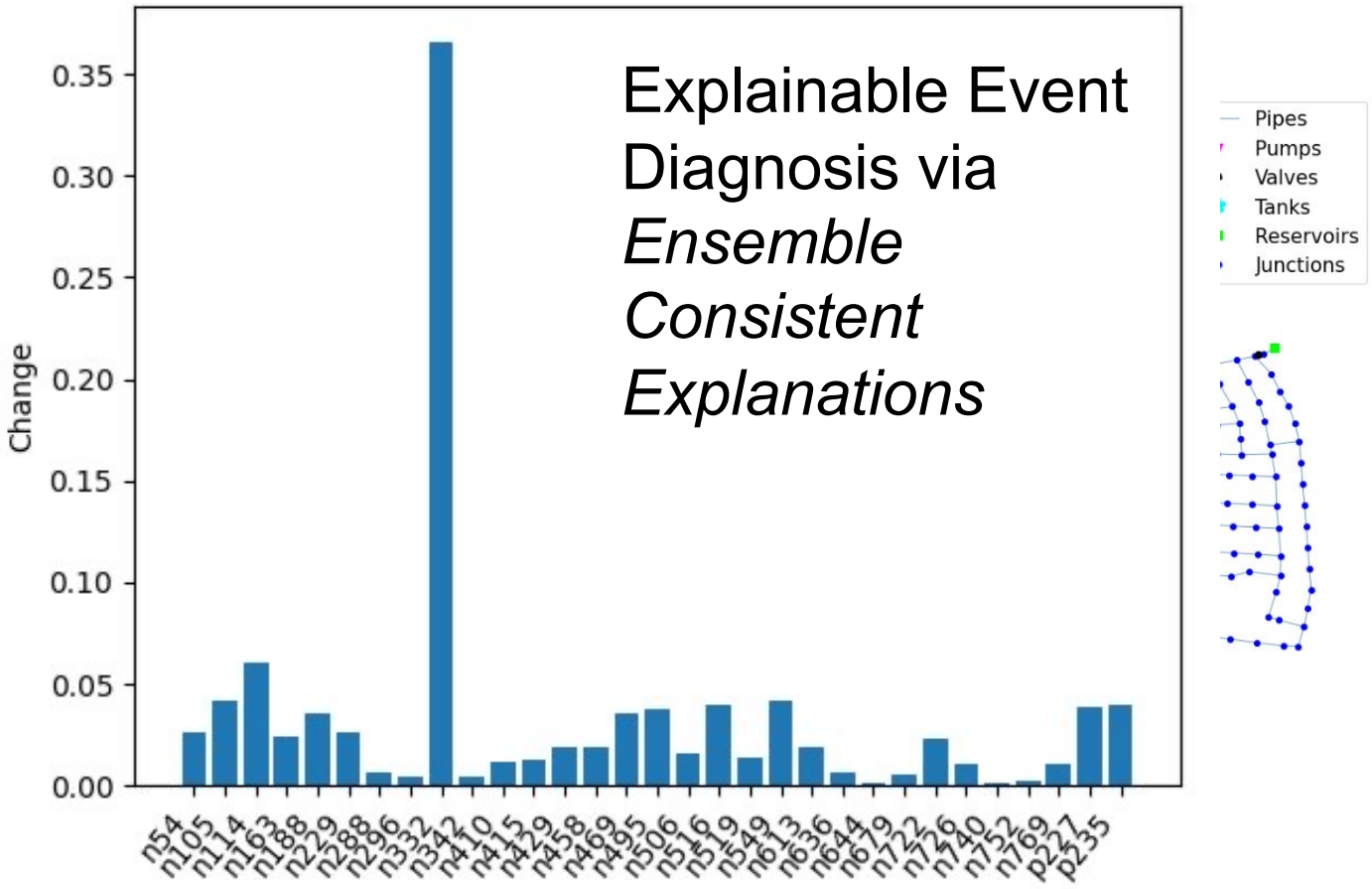
Counterfactual

New Prediction: 8

Kenny & Keane (2021) AAAI-21

Explaining Anomalies Across Virtual Sensors

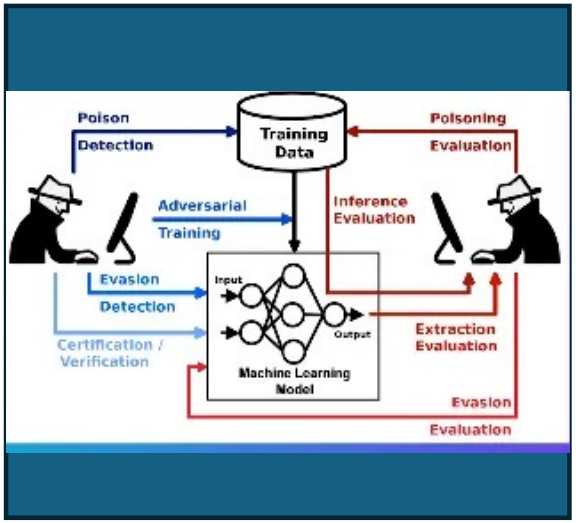
Counterfactual Explanations



Artelt et al. 2022 XAI Workshop @ IJCAI

Why Counterfactuals?

Artificial Intelligence



Looks like adversarial learning…

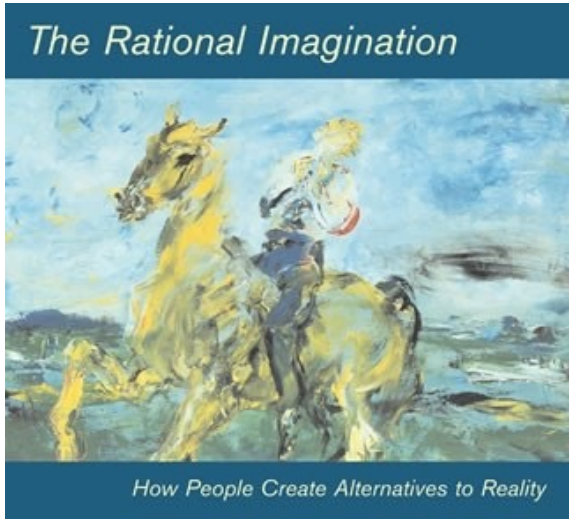
Wachter et al (2019) Facct-19

Legal

GDPR compliant…

Wachter et al (2017) Harv. JL &B Tech.

Psychology



Used by people spontaneously…

Byrne (2007) MIT Press

Plan for the Day...

CF Tutorial

TIME	Topics
9:00 AM	***Introduction***
	Hello and Introducing Ourselves!
	Hands-on: *Trying Our Study (follow link)*
9:30 AM	***Historical Fundamentals of Counterfactuals***
	From Philosophy to XAI (via Psychology)
	Two Sample User Studies and Q&A
10:30 AM	COFFEE (10:30-11:00)
11:00 AM	***Fundamentals of Counterfactuals in AI***
	Formalisation
	Modelling Approaches & Key Constraints
11:30 AM	***Using Counterfactual Algorithms***
	Hands-on: *A Counterfactual Toolbox (AA)*
	Hands-on: *Checking Out Notebooks and Q&A*
12:00 PM	***Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design
12:30 PM	LUNCH (12:30-14:00)
2:00 PM	***Algorithmic Growth Points***
	Computational Future Directions and Q&A
2:30 PM	***More Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design (cont.)
3:00 PM	COFFEE (15:00-15:30)
3:30 PM	***From Fundamentals to an Actual User Study***
	User Studies II: A More Complex Design
	User Studies III: Even More Complex Designs
	Hands-on: *Looking At Our Study*
5:00 PM	***Closing Session, Discussion and Final Q&A***
	TUTORIAL END

The Full Experience:
Be a Participant!

https://tinyurl.com/ijcai24-XAI